A MULTIVARIATE STUDY OF T/V FORMS IN EUROPEAN LANGUAGES BASED ON A PARALLEL CORPUS OF FILM SUBTITLES*

NATALIA LEVSHINA
Leipzig University
natalevs@gmail.com

Abstract
The present study investigates the cross-linguistic differences in the use of so-called T/V forms (e.g. French *tu* and *vous*, German *du* and *Sie*, Russian *ty* and *vy*) in ten European languages from different language families and genera. These constraints represent an elusive object of investigation because they depend on a large number of subtle contextual features and social distinctions, which should be cross-linguistically matched. Film subtitles in different languages offer a convenient solution because the situations of communication between film characters can serve as comparative concepts. I selected more than two hundred contexts that contain the pronouns *you* and *yourself* in the original English versions, which are then coded for fifteen contextual variables that describe the Speaker and the Hearer, their relationships and different situational properties. The creators of subtitles in the other languages have to choose between T and V when translating from English, where the T/V distinction is not expressed grammatically. On the basis of these situations translated in ten languages, I perform multivariate analyses using the method of conditional inference trees in order to identify the most relevant contextual variables that constrain the T/V variation in each language.

Keywords: T/V pronouns, politeness, film subtitles, conditional inference trees

1. Aims and challenges of this study

The present study investigates the cross-linguistic differences in the use of the so-called T/V forms (e.g. French *tu* and *vous*, German *du* and *Sie*, Russian *ty* and *vy*, usually accompanied by a corresponding verb form). The use of T/V, alongside with titles, names and other forms of address, often becomes a matter of public metalinguistic reflection and debate (e.g. Szarkowska 2013: 36–39). It also has great relevance for intercultural communication, translation, product localization and other practical purposes. However, although the grammatical forms representing this distinction are well described, we still know little about

* This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n° 670985). The author is sincerely grateful to the anonymous reviewers for their constructive suggestions. All remaning errors are solely mine.

the cross-linguistic similarities and differences regarding the communicative situations in which T and V are preferred. The previous empirical studies based on questionnaires and focus group interviews are illuminating, but rather limited in scope with regard to the number of languages (see Section 2).

This study aims to fill in this gap. It is based on the data from ten diverse European languages. Most of them are Indo-European and represent different genera: Germanic (Dutch, German, Swedish), Romance (French and Spanish), Slavic (Bulgarian, Polish, Russian), and Greek. One language, Finnish, belongs to the Finno-Ugric family. Although the sample does not include all European languages, I believe it is representative enough to give an idea of the magnitude of variation of T/V in Europe and to allow for cautious, yet evidence-based generalizations about the most important dimensions of this variation.

A possible reason of the lack of large-scaled comparative studies of the communicative constraints is that the latter represent an elusive object of investigation because they depend on a number of subtle contextual and social distinctions, which should be matched across the languages. Film subtitles in different languages offer a convenient solution because the situations of communication between film characters can serve as comparative concepts (Haspelmath 2010) and represent very diverse social situations and relationships. Thanks to the multimodal information available in films, one can analyse the communicative settings and evaluate the relationships between the speaker and the hearer, as well as their social class, age and other characteristics that may be relevant for the choice between T/V forms. Although there are some risks involved in using film subtitles for this kind of research, there is empirical evidence that subtitles represent spontaneous informal conversations quite faithfully (see Section 3.1).

The soundtracks of the films used as source texts in this case study were in English. At the present stage, this language does not have a T/V distinction, and the pronoun *you* is used in formal and informal situations regardless of social distance and power relationships.[1] In the process of translation, English forms of address undergo obligatory explicitation (Szarkowska 2013: 212), as the subtitle translators are forced to choose between T/V forms in the target language.

According to previous research, one can identify the prototypical social situations, where the choice is straightforward, and 'grey areas', where variation is possible (see Section 2.2). In this situation, it is only natural to investigate the constraints on the use of T/V with the help of quantitative methods, as it has been done in variational linguistics (e.g. Weiner and Labov 1983) and more recently in probabilistic grammar and Cognitive Sociolinguistics, where the constraints on the use of two or more linguistic variants are compared across language varieties (e.g. Szmrecsanyi et al. 2016). In this study, I employ a non-

[1] Sometimes the reflexive form *yourself* is preferred to indicate greater politeness and deference, e.g. *Is it for yourself?* [shop assistant addressing a customer who is considering a garment] (Carter & McMarthy 2006: 385).

parametric regression and classification method called conditional inference trees (cf. Tagliamonte and Baayen 2012) in order to compare the constraints across the languages.

The use of terms of address belongs to the vast body of research on politeness phenomena. Section 2 presents some important theories and concepts related to T/V forms and discusses the main parameters of variation in a selected sample of European languages. In Section 3, I describe my corpus of film subtitles and the data set, focusing also on the pros and cons of using film subtitles for the purposes of the present study. Section 4 contains a list of variables that were tested. In Section 5, I present the results of the multivariate analyses, which employ conditional inference trees in order to compare the constraints cross-linguistically. Finally, Section 6 summarizes the results and suggests some directions for future research.

2. Previous research

This section presents the main sociolinguistic dimensions of the use of T/V forms in European languages (Section 2.1) and provides a brief summary of previous findings related to the cross-linguistic differences in the constraints (Section 2.2).

2.1. Common dimensions: power and solidarity

The foundations of analysis of T/V forms were laid in a highly influential paper by Brown and Gilman (1960). As they point out, the T/V variation existed already in Old French, Old Italian, Old Spanish, Old Portuguese and Middle English, although it is very difficult to pinpoint the rules for that period. However, between the 12th-14th centuries, the set of norms crystallized, which can be described as nonreciprocal power semantics (*Ibid.*). Power means the ability of one person to control the behaviour of another one. In a one-to-one interaction, the participant with greater power addressed the participant with less power using T, whereas the participant with less power used V when addressing a more powerful person. Examples of such power dyads are a father and his child, a king and his vassal, a nobleman and his servant, a priest and a penitent. These relationships are asymmetric and vertical. Gradually, however, these uses became associated with entire classes and social groups. For example, in the 17th century France, noblemen and merchants always addressed people from their social class using V, which was considered a mark of elegance. People from lower classes, such as servants and peasants, always used T (*Ibid*: 257).

During the 19th century, the power system was gradually replaced by reciprocal solidarity semantics, with T for intimate communication, e.g. between family members and friends, and V for formal communication. These relationships are symmetric and horizontal. Virtually any dimension, e.g. gender,

age, the school one went to, political persuasions and hobbies can be basis for the perception of solidarity or non-solidarity. This can even include physical appearance, e.g. if both the speaker and the hearer wear dreadlocks, T would be preferred (Warren, 2006).

However, V forms can be perceived not only as distant, but also as respectful. This is a manifestation of negative politeness, which reflects the desire of the speaker to avoid imposition on the hearer. It contrasts with positive politeness, which is associated with appreciation, consideration and solidarity (Brown and Levinson, 1987). At the same time, T forms may be perceived either as warm and friendly (a manifestation of positive politeness), or as too familiar (from the perspective of negative politeness). For example, Wierzbicka (1985: 171) argues that the Polish address system expresses cultural values of intimacy and courtesy, where T form *ty* is intimate and V form *Pan/Pani* is "courteous and personal" and "based on mutual respect". Similarly, although for some French speakers *vous* represents distance, for others it is associated with respect (Warren 2006). Thus, the reciprocal semantics in fact has two sides, which correspond to positive and negative politeness, or solidarity and respect.

2.2. Cross-linguistic variation in the use of T and V

From a typological perspective, about one quarter of languages in the world have a politeness distinction in the pronouns (Hembrecht, 2012). In North and South America, New Guinea, Australia, and most of Africa there are no politeness distinctions in personal pronouns. The hotbed of pronouns with a binary distinction is Europe and adjacent areas, although they also occur in other regions (*Ibid.*). Thus, we can regard the use of T/V pronouns as an areal phenomenon. The diachronic and synchronic similarities between the European languages, which were pinned down by Brown and Gilman (see Section 2.1) can be explained by extensive cultural and linguistic contact between the European countries during many centuries.

However, there are also cross-linguistic differences in the use of T/V forms, as one can conclude from numerous studies describing the forms of address in European languages within the general framework of politeness research (e.g. Hickey and Stewart, 2005), although most of them are qualitative and focus on one language and a set of particular situations (e.g. service encounters, interviews, online forums). A remarkable exception is the comparative project 'Address in Some Western European Languages' (cf. Kretzenbacher et al., 2006; Norbby, 2006; Warren, 2006), where the researchers used interviews of focus groups and some other methods to elicit the use of T/V in three European languages (French, German and Swedish) in different locations.

Generally speaking, one can propose a preliminary typology of European languages with regard to T/V distinction. First, there are languages where both T and V are used more or less on a par. Examples of such languages are French, German and Russian. Second, there are mostly T languages, where the V form is

marginal and is used in very specific contexts. Examples of such languages are Swedish and Finnish. Languages from the third group are in-between (e.g. present-day Italian, cf. Molinelli, 2015: 290).

The speakers of T/V languages know well the prototypical situations in which only one form is appropriate. The prototype of T usually includes communication among family members and close friends. It is also frequently used among younger people and informal contexts. The prototypical uses of V forms normally include addressing strangers, official contexts (authorities) and service encounters (e.g. Kretzenbacher et al., 2006; Warren, 2006). The cross-linguistic and intra-linguistic variation mainly concerns the grey zone between these prototypes.

Overall, the use of the forms is reciprocal and based on solidarity. The greater the perceived similarity between the communicators, the greater the chances of T being used. However, the specific features and dimensions of this similarity may differ cross-culturally and even individually and situationally. For instance, according to Brown and Gilman's (1960) pioneering study based on the questionnaire data from male upper-class European students in American universities, the solidarity expressed by T used by the German students was based on family relationships, whereas the solidarity for the French students depended more on sharing a common life story. The Italian students used T more frequently than both the French and German students, and the solidarity was also extended to the female fellow students.

According to a more recent and inclusive study based on interviews of focus groups, the French often use T between people of the same sex (Warren, 2006), whereas German speakers may pay attention to the relative age, emotional closeness, commonalities in lifestyle and length of co-residence (e.g. with neighbours) (Kretzenbacher et al., 2006). At the same time, the construal of distance may also depend on political views of the speaker. For example, a German leftist will also use the T form *du* more frequently in all situations than people with other political views (*Ibid.*).

Perceived similarity may also be situational. For example, Friedman (1972: 276) gives an example from a Russian novel, when two officers exchange *vy* when discussing military tactics, but switch to *ty* back in their quarters when they chat about women. Irony and sarcasm can also be expressed by using a V form when a T form is appropriate. There is substantial variation and room for individual preferences and negotiation (Warren, 2006). On a less positive side, it appears that T/V forms have a high "embarrassment potential" (Kretzenbacher et al., 2006).

Swedish represents a particularly interesting case. The use of the Swedish 2nd person pronouns has been substantially influenced by a national language policy. Unlike in other languages, the system of address in Swedish underwent a radical transformation in the 1960–1970s, which was initiated by media and intellectuals and has remained in history as the *du-reformen*. The previous complex system of V forms with titles used towards superiors and *ni* used

towards inferiors meant that there was a lacuna in the system: there was no neutral, polite form of address. Although there were some attempts to re-introduce *ni* in that function, the egalitarian ideas in the 1960s and 1970s resulted in a wide spread of the T pronoun *du* as a democratic, no-nonsense form (Norbby, 2006). At the moment, the recent Norrby's (2006) interviews of focus groups demonstrate that the V pronoun *ni* is gradually disappearing from Swedish. The T pronoun *du* is the default form in most situations. Some variation is possible when speaking to old people and strangers. In Finnish, the situation is rather similar, and the T pronoun *sinä* is used in most cases.

Overall, many researchers observe that T forms are gradually winning the territory also in T/V languages. However, there is a counter-trend, and the younger generation may be more conservative in that respect than the generation of the 1960s. For example, Swedish experiences a re-entry of *ni* in service encounters between the young and the middle-aged and old speakers. In Finnish, old and some young people find the V pronoun *te* to be more acceptable in communication between strangers (e.g. service encounters) than the "Beatles generation", who usually prefer the T form *sinä* (Yli-Vakkuri, 2005). The reluctance to use the T form as the default in Finnish has also resulted in emergence of a subtle system of avoidance of direct address, such as in impersonal constructions (*Ibid*.). All these developments can be seen as a manifestation of the increasing importance of negative politeness.

In addition to all that, many languages exhibit substantial regional variation. For example, the V form *ni* is more frequently used in Finnish Swedish than in Sweden (Norbby, 2006). As for the German-speaking area, the T pronoun *du* between colleagues is more common in Vienna (Austria) than in Mannheim (West Germany). In Leipzig (East Germany), it is the least common. A possible explanation may be a lasting reaction to the GDR ideology (Kretzenbacher et al., 2006).

The aim of the present study is to pinpoint the differences on the basis of corpus evidence and to test and extend the results of the previous studies. The data source and variables are described below.

3. Data from a parallel corpus of film subtitles

3.1. Why film subtitles?

The nature of subtitle translations is very different from that of natural conversations. In the latter, the use of T/V forms and other politeness markers reflects the Speaker's construal of the communicative situation and the relationships with the Hearer in terms of social distance, power and other parameters (see Section 2.1). In film subtitles, the cognitive mechanisms involved are much more complex. These mechanisms are outlined below.

(1) The first step where cognitive construals are involved is the creation of a script by the script writers. The film dialogue thus represents the authors' idea of natural dialogue. Moreover, according to Bell's theory of audience design (Bell 1984), it reflects their idea of the future audience, who are the likely receivers of the message, and how the latter will interpret the interaction shown on the screen (Hatim and Mason, 1997: 82–84). Later, the actors may change the script when the film is shot, adding their own vision of what the dialogue should sound like in a particular situation. According to previous research into transcribed TV series dialogues and films, one of the most striking differences between natural and scripted dialogue is the lower frequency of narrative and 'vague' elements and some discourse markers in the latter (e.g. Mittmann, 2006; Quaglio, 2008; Bednarek, 2011). At the same time, many researchers observe substantial similarity between these two types of dialogue regarding various lexico-grammatical and pragmatic features (see Dose, 2014: Ch. 4.3.4 for an overview).

(2) Next, the translators as viewers should interpret the communicative situations shown in a film in order

> to make conscious decisions about the nature of the relationships among different characters in the story and about the social standing of these characters as reflected in their adoption of certain conventions to do with approved/non-approved expression of familiarity and/or deference. (Baker 1992: 97)

This is not an easy task, which requires extensive background knowledge of the culture and the plot. For example, the Slovak translator of the pilot series of *House M.D.* uses a T form in the communication between Dr. House and his boss because, as shown in the following episodes, they attended the same university and had an affair. In contrast, the Czech translator uses V, obviously, not aware of their previous relationships, which are revealed only later (Marketa Janebova, p.c.).

(3) After that, the translators should encode these relationships in the target language (cf. Odber de Baubeta, 1992). When doing this, they should follow the probabilistic rules of using T/V forms in the target language. One should also be aware of potential individual variation between translators, who may belong to different social classes and demographic groups, and may have somewhat different mental representations of the constraints on the use of T/V in their language, as pointed out by Braun (1988: 24ff). This is why it is desirable to collect data from many different translators.

(4) Finally, the process of creating subtitles has its own rules and limitations. In particular, professional subtitlers follow strict rules with regard to the number of characters per line, number of lines, duration of a subtitle or caption on screen, etc. (e.g. Díaz Cintas and Remael, 2007[2014]: Ch. 4; Deckert, 2013: App. 1). Although the online film subtitles used in this study are mostly created by amateur translators (see Section 3), these limitations are still very important.

All these *a priori* considerations do not make online film subtitles the most obvious choice. However, there is empirical evidence that film subtitles can represent the target language quite faithfully. First, previous psycholinguistic research has demonstrated that film subtitles are a reliable source of lexical norms and that subtitles sometimes even outperform other sources (Keuleers et al., 2010). Moreover, a quantitative study based on *n*-gram frequencies demonstrates that English subtitles, original and translated from other languages, are highly similar to informal spontaneous conversations from such well-established sources as the British National Corpus and Santa Barbara Corpus of Spoken American English (Levshina, Forthc.). There is also some support from translation studies. In particular, Szarkowska (2013: 138), who investigates English subtitles of Polish TV soap operas, observes that subtitlers tend to adhere to the norms of the target language when translating different terms of address.

3.2. The corpus and the procedure of data extraction

The data for the present study come from online subtitles of nine popular films of different genres. The subtitles were downloaded from the website www.opensubtitles.org and constitute part of the ParTy corpus (Parallel corpus for Typology) (Levshina 2016). The films are displayed in Table 1. The meta-information about the year and genres is taken from the International Movies Database.[2]

Table 1. Films represented in the data set

Film	Year	Genres
Avatar	2009	Action, adventure, fantasy
Black Swan	2010	Drama, thriller
Bridge of Spies	2015	Drama, history, thriller
Frozen	2013	Animation, adventure, comedy
Inception	2010	Action, adventure, sci-fi
Spectre	2015	Action, adventure, thriller
The Grand Budapest Hotel	2014	Adventure, comedy, crime
The Imitation Game	2014	Biography, drama, thriller
The Iron Lady	2011	Biography, drama, history

Naturally, no one can guarantee the quality of film subtitles downloaded from online repositories. There is hardly any reliable information about the subtitler of a specific film and his/her linguistic background and expertise. However, my personal experience based on linguistic analysis of subtitles and native speakers'

[2] www.imdb.com, last access 25.08.2016.

evaluation of text samples in several languages, suggests that the vast majority of subtitles are close to natural informal discourse (although one can see occasional mistakes). In addition, many subtitles undergo several rounds of corrections from online comments or meta-information in the files.

All subtitles were aligned with the English version. The data set for the study was created as follows. First, I identified 243 contexts where the pronouns *you* or *yourself* were used in the English version of the subtitles. Plural reference was excluded. The multimodal data allowed me to distinguish between singular and plural addressees seamlessly.

Next, I identified the personal forms used in the translations and coded them as T or V. An overview of the T/V forms in the ten languages is provided in Table 2. The Polish V pronouns *pan* (m) and *pani* (f), which have a gender distinction, are unique among the European languages because of their double function: they are homonymous with the nouns that represent titles, such as *Mr./Mrs*. or *Herr/ Frau* (Łaziński 2006).

Table 2. T and V forms in the languages represented in this case study

Language	T pronoun	V pronoun	V verb agreement
Bulgarian	ти [ti]	Вие [ˈvi.ɛ]	2nd PL
Dutch	jij (je)	u	2nd person SG
Finnish	sinä	te	2nd PL
French	tu	vous	2nd PL
German	du	Sie	3rd person PL
Greek	εσύ [eˈsi]	εσείς [eˈsis]	2nd PL
Polish	ty	pan (m)/pani (f)	3rd person SG
Russian	ты [ty]	Вы [vy]	2nd PL
Spanish	tú	usted	3rd person SG
Swedish	du	ni	2nd PL

It is important to mention that one should speak about T/V forms, rather than about T/V pronouns in this context. The reason is that this list includes subject pro-drop languages (cf. Dryer 2013), where the subject is usually left unexpressed (Bulgarian, Finnish,[3] Greek, Polish, Spanish), as in the Polish translation (1).

[3] Finnish belongs to the 'mixed' type, i.e. the first and second person subject pronouns are usually absent, and the third person pronouns are obligatory (Dryer 2013).

(1)

English original: *What are you doing here?* (*The Grand Hotel Budapest*)
Polish: *Co tu robisz?*
what here do.IPF.PRS.2SG
"What are you doing here?"

When the personal pronoun is left out, the verb form becomes the only indicator of the choice between T and V. A similar case is the omission of pronouns in the imperative.

The frequencies of different translational options are displayed in Table 3.

Table 3. Frequencies of T, V and other forms found in the translations

Language	**T**	**V**	**Other forms**	**No personal forms**
Bulgarian	145	86	0	12
Dutch	159	73	0	11
Finnish	145	72	1	25
French	71	152	0	20
German	91	133	6	13
Greek	160	76	0	7
Polish	150	70	3	20
Russian	112	120	0	11
Spanish	141	93	1	8
Swedish	159	66	0	18

The table shows that the lowest number of T forms in the translations of *you* and *yourself* are observed in French, followed by German and Russian. In these languages, the V form is the more frequent one. The highest frequency of the T form is observed in Greek, followed very closely by Swedish and Dutch. In Swedish, the frequency of V is the lowest. Notably, the highest frequency of absence of any personal form is found in Finnish (10% of all examples). This finding confirms the results of previous research of avoidance strategies in that language (see Section 2.2). However, very commonly the lack of personal forms is due to the differences in formulaic language, e.g. in (2):

(2)

English original: *I thank you, sir.* (*Bridge of Spies*)
Dutch: *Bedankt, sir.*
French: *Merci, monsieur.*
Spanish: *Gracias, señor*.

The forms under the category 'Other' are very infrequent. In German, where this category has the highest frequency (i.e. six times), the 2nd person plural forms with pronoun *Ihr* represent an archaic formal form of address. Consider an example in (3):

(3)

English original: *Sorry to wake you, ma'am.* (*Frozen*)
German: *Verzeiht* *mir.*
forgive.IMP.2PL me.DAT
"I beg your pardon."

The next step is to investigate and compare the constraints on the use of T/V across the languages. For that purpose, the data were coded for several variables based on an in-depth analysis of multimodal information from the films. These variables are described in Section 4.

4. Contextual variables

The variables can be subdivided into four types: relational (representing relationships between the Speaker and the Hearer in a dyad), characterizing the Speaker only, characterizing the Hearer only and describing the communication settings.

4.1. Relational variables

These variables describe the potential differences or similarities in the social and demographic characteristics of the Speaker and Hearer in the dyad.

- *Rel_Age*: whether the Hearer is older or younger than the Speaker, with the values "Same", "Older" or "Younger".
- *Rel_Power:* whether there is power asymmetry between the participants in general or in the given situation. Examples of power relationships are the relationships between a general and a soldier, a boss and an employee, a parent and a child, or a terrorist and a hostage. The values are "Greater" (the Hearer has power over the Speaker), "Less" (the Speaker has power over the Hearer) or "Equal";
- *Rel_Class:* the social class difference in the dyad. The values are "Higher" (the Hearer belongs to a higher social class than the Speaker), "Lower" (the Hearer belongs to a lower social class than the Hearer) or "Equal";
- *Rel_Gender:* the gender of the Speaker and the Hearer, with the values "F_F" (female Speaker and female Hearer), "F_M" (female Speaker and male Hearer), "M_F" (male Speaker and female Hearer) and "M_M" (male Speaker and male Hearer);

- *Rel_Circle:* the social circle to which the Speaker and the Hearer belong. The values are "Fam" (family), "Fri" (friends), "Rom" (romantic partners), "Home" (unrelated people living in the same place, e.g. servants in a household or guests at a hotel), "Pri" (prison), "Aca" (school, university), "Work", "Acq" (acquaintances) and "Str" (strangers).

4.2. Speaker-related variables

These variables describe the characteristics of the Speaker only.

- *S_Age*, the speaker's age: "Child" (younger than 18), "Young" (approximately 18–35), "Middle" (approximately 35–60), "Old" (approximately older than 60);
- *S_Class*, the speaker's social class: "Upper" (top-rank politicians and civil servants, owners of multinational corporations, etc.), "Middle" (white-collar workers, small business owners, military officers, etc.), "Lower" (blue-collar workers, servants, etc.) and "Other" (aliens, animals, as well as gangsters, tramps, prostitutes and other declassed elements);
- *S_Gender*, the speaker's gender: "M" or "F" (there were no transgenders in the data).

4.3. Hearer-related variables

These variables describe the characteristics of the Hearer only. They mirror the ones related to the speaker: *H_Age*, *H_Class* and *H_Gender*.

4.4. Variables describing the communicative settings

- *Others*, the presence of other people who could hear the Speaker, with the values "Yes or No";
- *Office*, whether the interaction takes place in an office, a government building, prison, school, etc. ("Yes") or in another place, e.g. in a bar or in the street ("No");
- *Before68*, whether the action takes place before 1968, with the values "Yes" or "No". This year is selected as a cut-off point because it was the time of profound social and cultural changes towards a more democratic and equal society;
- *Britain*, whether the action takes place in the United Kingdom or not, "Yes" or "No". This variable was added because the British culture is often perceived as more formal than the other English-speaking cultures, especially the American one.

5. Quantitative analyses: conditional inference trees

This study employs a non-parametric regression and classification technique, which is called conditional inference trees (Hothorn et al., 2006). This method is appropriate in the situations when the number of predictors is large relative to the number of observations and therefore the data are not appropriate for traditional logistic regression. This is exactly the case in this study. There are fifteen contextual variables and only 243 observations. This method can also deal with highly correlated predictor variables and complex interactions. It is easy to see that many of the variables are strongly associated, e.g. the gender of the Speaker and the Hearer taken separately (*S_Gender* and *H_Gender*) and together (*Rel_Gender*). In linguistics, conditional inference trees and random forests, which are 'grown' from many trees, have been applied in variational studies (Tagliamonte and Baayen 2012), comparative corpus-based studies of morphosyntactic phenomena (Levshina 2016) and in some other domains. The approach used here is superior to other classification and regression methods, such as the traditional CART algorithm, because the trees do not have to be pruned (Hothorn et al., 2006). That is, the algorithm 'knows' where to stop adding new branches to the tree. This method is also unbiased with regard to the number of categories in a categorical variable. This approach is implemented in the package *party* (Strobl et al., 2008) in R (R Core Team, 2015).

As an illustration, consider Figure 1.[4] It displays a conditional inference tree based on the Dutch data. The figure can be interpreted in the following way. In the beginning, the algorithm takes all data available for Dutch (excluding the observations without personal forms) and finds the contextual variable that is associated the most strongly with the response, i.e. the use of T/V forms. This is *Rel_Class*, shown in Node 1. The permutation-based *p*-value is less than 0.001, which suggests that this association is highly significant. The data are then split into two subsets represented by two branches. One (the left branch) contains all observations with the value *Rel_Class* = "Higher" and the other one (the right branch) with all other values ("Equal", "Lower" and "Other"). The procedure is then repeated on both subsets. If the value is "Higher", the algorithm does not find another variable which would be associated with the response at the level of significance of 0.05. The final node (Node 2) contains 42 observations. The distribution of the T/V forms is represented by a bar chart. The darker shading represents the V form *u*, and the lighter shading represents the T form *jij*. The relative sizes of the differently coloured areas show that *u* is used in the vast majority of cases when *Rel_Class* = "Higher".

[4] The tree in Figure 1 may resemble the famous flow-chart in Ervin-Tripp (1972), where she shows the rules that determine the choice between particular forms of address in American English. Indeed, the convergence is remarkable. However, the algorithm presented here is probabilistic, automatic and data-driven.

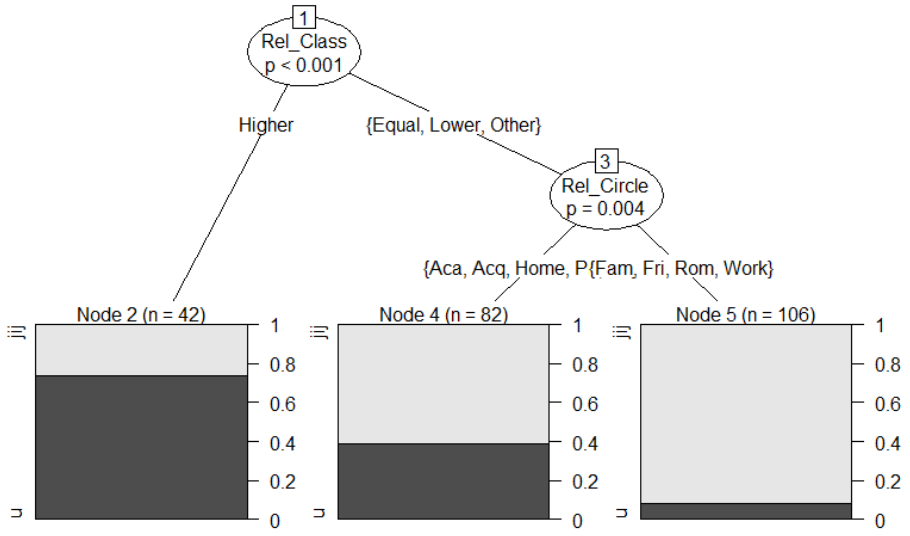

Figure 1. Conditional inference tree for Dutch

Now let us explore the right branch, which corresponds to the observations with all other values of *Rel_Class*. The next split is made in Node 3. It is the variable *Rel_Circle*, which is associated with the response significantly, $p = 0.004$. The split separates the categories "Aca" (school and university), "Acq" (acquaintances), "Home" (people living in the same house), "Pri" (prison) and "Str" (strangers), from the categories "Fam" (family), "Fri" (friends), "Rom" (romantic partners) and "Work" (work colleagues). Note that some of the labels are masked due to overplotting. The first group, represented by the left branch, contains 82 observations. The second group, represented by the right branch, totals 106 observations. As the bar charts in Nodes 4 and 5 reveal, the most common choice in both types of situations is the T form *jij*. However, the observations in the left branch contain a slightly higher proportion of *u* than the ones in the right branch. This difference indicates that *u* is more acceptable when the relationship is less intimate. Although the difference between the proportions of *u* and *jij* in Node 4 and Node 5 is small, it is still statistically significant at the conventional level of significance ($p = 0.004$). The absence of further splits suggests that no more contextual variables are associated with the response under the criteria specified above.

The results of an examination of all ten conditional inference trees are described below variable by variable.

1) *Rel_Age*. This variable appears only in the Polish tree. It splits the observations where the Hearer is older than the Speaker from all others. The observations with older Hearers contain significantly more V forms than the

other observations. However, this variable only matters when the interlocutors are strangers, acquaintances or prison mates, and when addressing the Hearer whose social class is not higher than that of the Speaker.

2) *Rel_Power*. This variable did not appear in any tree.

3) *Rel_Class*. This variable is present in many trees. In Bulgarian, Dutch, Finnish, Greek and Polish, the Hearer who belongs to a higher social class than the Speaker, is called more frequently by the V form. In Bulgarian and Finnish, this variable is relevant only for the high-intimacy social circle (including friends, family, colleagues and people at the same school or university), and irrelevant for more distant relationships. In Russian, the effect is different. The declassed elements and non-human beings (*Rel_Class* = "Other") get significantly more T forms than the others, although this holds only for the low-intimacy social circles and among younger speakers, children and, again, non-humans. This effect seems to be due mostly to the situations involving aliens, magic beings and animals. In addition, this variable is also important in Swedish, but only in the interactions that are assumed to have happened before 1968. If the social class of the Hearer and the Speaker is not equal, one observes significantly more V forms.

4) *Rel_Gender*. This variable appears only in the Finnish tree. Men speaking to male strangers or acquaintances use significantly more V forms than men speaking to women, women speaking to men or women speaking to women in similar situations.

5) *Rel_Circle*. This variable is important in all languages, except Swedish. In all these languages, communication between strangers and acquaintances is associated with more V forms than that between family members and friends. However, there is some variation in the rarer categories (school or university, prison, romantic partners, household members). These cross-linguistic differences are summarized in Table 4. The table lists only the social circles that favour T forms. The social circles that favour the use of V forms are those that are not mentioned in the table.

Table 4. Effects of the social circle (variable *Rel_Circle*) and the conditions when the split is made

Language	T-forms	Conditions of the split
Bulgarian	school or university, family, friends, prison, romance, work	None (applicable to all situations)
Dutch	family, friends, romance, work	only if the Hearer doesn't belong to a higher social class than the Speaker
Finnish	school or university, family, friends, prison, romance, work	None (applicable to all situations)
French	family, friends	No (applicable to all situations)
German	family, friends	only if the Hearer is middle-aged or old
Greek	school or university, family, friends, living in the same place, prison, romance, work	only if the Hearer doesn't belong to a higher social class than the Speaker

Polish	school or university, family, friends, living in the same place, romance, work	if the Hearer doesn't belong to a higher social class than the Speaker
Russian	school or university, family, friends, prison, romance	None (applicable to all situations)
Swedish	No significant effect	NA
Spanish	family, friends, romance	None (applicable to all situations)

In addition, in Dutch, German, Greek and Polish, the social circle has a significant effect only if the social class of the Hearer is not higher than that of the Speaker (see the column "Conditions of the split"). This set suggests that power semantics sometimes override solidarity semantics. In Swedish, the social circle does not have a significant effect.

6) Speaker-related variables. From all these variables, only *S_Age* has been found in the Bulgarian tree. If the Speaker is young, a child or is a non-human, one can expect significantly more T forms. However, this variable only matters in low-intimacy circles (acquaintances, strangers, unrelated people living in one home).

7) Hearer-related variables. Only the variable *H_Age* appears in the trees. It does so in five languages (French, German, Greek, Russian, Spanish). In all of them, the Hearers who are children, young people or non-human beings, are called significantly more frequently by using T forms. This variable, however, is only relevant for the low-intimacy social circles in French, Russian, Greek and Spanish. In Greek, there are additional conditions. The Hearer's age matters only when the Hearer does not belong to a higher social class than the Speaker and the interaction does not happen in Britain.

8) *Others*. The presence or absence of others does not play a significant role in any of the languages.

9) *Office*. This variable matters only in German, where interactions that take place in a public place are associated with significantly more V forms. This distinction is relevant only for Hearers who are young or children. Apparently, young people tend to be addressed as V in the office and as T outside.

10) *Before68*: The interactions that took place before 1968 are associated with more V forms in Swedish and French. In French, this distinction is important only for younger Speakers who are not family or friends.

11) *Britain*. If the interaction takes place in Britain, one can expect significantly more V forms in Bulgarian, Greek, Polish, Spanish. However, the effect is rather marginal and depends on numerous conditions in the situations that often include equal and/or young strangers.

6. Conclusions and outlook

The quantitative analyses have revealed substantial cross-linguistic variation, regarding both the quantitative and qualitative aspects of T/V use. The results also allow us to make a number of generalizations.

First, in accordance with the Brown and Gilman's (1960) theory, the solidarity dimension, which is represented by the social circle, plays a crucial role in all languages, with the exception of Swedish, where the proportion of V forms is the lowest. This is the parameter where the languages agree the most. However, although the prototypical situations where T and V are used, are the same (family and friends vs. strangers and acquaintances), there is substantial cross-linguistic variation in the 'grey zone'. T forms are the most restricted socially in French and German (including only friends and family), and the least restricted in Greek, Finnish and the Slavic languages. In addition, the place of communication (office or not) matters for young Hearers in German.

Yet, power semantics is still present in quite a few languages. It is expressed in the situations when the Speaker prefers V form when addressing the Hearer who belongs to a higher social class. This effect is found in Bulgarian, Dutch, Finnish, Greek and Polish. Not surprisingly, this parameter is also important in communication between non-equals in Swedish before 1968. However, the other manifestations of power relationships (based on relative age, gender difference, power asymmetry) are not important, with the exception of the relative age in Polish in a limited number of situations. This is another common feature shared by the languages.

Interestingly, the individual characteristics of the Hearer, in particular, the age, are more important than those of the Speaker. For instance, one commonly uses T when addressing children and young people. This constraint is observed in many languages. The Speaker's age is only important in one language (Bulgarian).

An intriguing finding comes from Finnish, where the gender seems to play a role. Namely, communication between men who do not know each other well is more associated with V forms than in the situations when women are involved. It may be a manifestation of negative politeness between equals, which may be more relevant for Finnish men.

Based on these observations, one can propose the following general scale:

Solidarity > Power (social class) > Hearer only (age) > Other

where the cross-linguistic predictive power of the dimensions or groups of variables decreases from left to right.

The results also suggest that the translators have a sociolinguistic model not only of the present-day communication, but also of the norms of the past. This is why the time variable (whether interaction takes place before or after 1968) is

important in French and Swedish. In these countries, the 1960s left a particularly profound mark on the society.

In her cross-linguistic study of different address systems, Braun expressed scepticism towards Brown and Gilman's neat two-dimensional account: "Dealing with a number of address systems, however, doubt arises whether variants of address (…) indeed constitute such a common hierarchy and operate on two dimensions only" (Braun, 1988: 38). We can conclude, on the basis of the results of the present quantitative study, that Braun's scepticism is justified. Indeed, instead of the neat two dimensions, we have a complex picture of multifactorial and probabilistic variation. At the same time, it is impossible to deny that Brown and Gilman's pioneering work contains many insights that are still highly relevant today.

There are quite a few tasks that could be carried out in the future. Obviously, increasing the sample of languages and analysing more films of different types would lead to new insights. The novel thought-provoking details that have been discovered, such as the role of gender in Finnish, can be used for the formulation of hypotheses about the use of T/V in these and other languages of the world and tested on the data from non-translated naturally occurring discourse.

Necessarily, this quantitative study focused on the general tendencies and zoomed out of the subtle contextual modulations and negotiation of face. Investigation of such phenomena requires a careful qualitative analysis, which is beyond the scope of this paper. The cross-linguistic differences in the use of various terms of address, such as first names, diminutives, patronymics, titles, etc., which undoubtedly play an important role in politeness strategies, remain a task for future research, as well.

Another question for the future is how to explain the observed variation. Possible explanations may involve historical events (e.g. the social changes in Sweden in the 1960s), language contact (e.g. the Russian politeness system was greatly influenced by the French one) or language-internal factors. One of possible hypotheses, for example, would be that omission of personal pronouns might decrease the face-threatening potential of T and therefore increase the acceptability of T forms. This and other hypotheses are left for future research.

References

Baker, Mona. 1992. *In Other Words – A Coursebook on Translation*. London: Routledge.

Bednarek, Monika. 2011, The Language of Fictional Television: A Case Study of the 'Dramedy' *Gilmore Girls*. *English Text Construction* 4(1). 54–84.

Bell, Alan. 1984. Language Style as Audience Design. *Language in Society* 13. 145–204.

Braun, Friederike. 1988. *Terms of Address: Problems of Patterns and Usage in Various Languages and Cultures*. Berlin: Mouton de Gruyter.

Brown, Penelope and Stephen C. Levinson. 1987. *Politeness. Some Universals in Language Usage*. Cambridge: Cambridge University Press.

Brown, Roger and Albert Gilman. 1960. The Pronouns of Power and Solidarity. In Thomas A. Sebeok. *Style in Language*, 253–276. Cambridge, MA: MIT Press.

Carter, Ronald and Michael McCarthy. 2006. *Cambridge Grammar of English*. Cambridge: Cambridge University Press.

Deckert, Mikołaj. 2013. *Meaning in Subtitling: Toward a Contrastive Cognitive Semantic Model*. Frankfurt am Main: Peter Lang.

Díaz Cintas, Jorge and Aline Remael. 2007[2014]. *Audiovisual Translation: Subtitling*. London/New York: Routledge.

Dose, Stefanie. 2014. *Describing and Teaching Spoken English: An Educational-Linguistic Study of Scripted Speech*. PhD Dissertation. Giessen: Justus-Liebig-Universität Giessen.

Dryer, Matthew S. 2013. Expression of Pronominal Subjects. In Matthew S. Dryer & Martin Haspelmath (eds.), *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available online at http://wals.info/chapter/101 [Accessed: 5 September 2016.]

Ervin-Tripp, Susan. 1972. On Sociolinguistic Rules: Alternation and Co-occurrence. In John Gumperz and Dell Hymes (eds.), *Directions in Sociolinguistics: The Ethnography of Communication*, 213–250. New York: Holt, Rinehart and Winston.

Friedrich, Paul. 1972. Social Context and Semantic Feature: The Russian Pronominal Usage. In John J. Gumperz and Dell Hymes (eds.), *Directions in Sociolinguistics: The Ethnography of Communication*, 270-300. New York: Holt, Rinehart and Winston.

Haspelmath, Martin. 2010. Comparative Concepts and Descriptive Categories in Crosslinguistic Studies. *Language* 86(3). 663–687.

Hatim, Basil and Ian Mason. 1997. *The Translator as Communicator*. London/New York: Routledge.

Hickley, Leo and Miranda Stewart (eds.). 2005. *Politeness in Europe*. Clevedon: Multilingual Matters.

Helmbrecht, Johannes. 2013. Politeness Distinctions in Pronouns. In Matthew S. Dryer and Martin Haspelmath (eds.), *The World Atlas of Language Structures Online*.
Leipzig: Max Planck Institute for Evolutionary Anthropology.
Available from: http://wals.info/chapter/45 [Accessed: 3 August 2016].

Hothorn, Torsten, Kurt Hornik and Achim Zeileis. 2006. Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics* 15(3). 651–674.

Keuleers, Emmanuel, Marc Brysbaert and Boris New. 2010. SUBTLEX-NL: A New Frequency Measure for Dutch Words Based on Film Subtitles. *Behavior Research Methods* 42. 643–650.

Kretzenbacher, Heinz L., Michael Clyne and Doris Schüpbach. 2006. Pronominal Address in German: Rules, Anarchy and Embarrassment Potential. *Australian Review of Applied Linguistics* 29(2). 17.1–17.8.

Łaziński, Marek. 2006. *O Panach i Paniach. Polskie rzeczowniki tytularne i ich asymetria rodzajowo-płciowa*. Warszawa: Wydawnictwo Naukowe PWN.

Levshina, Natalia. Forthcoming. Online Film Subtitles as a Corpus: an N-Gram Approach. To appear in *Corpora.*

Levshina, Natalia. 2016. Why We Need a Token-Based Typology: a Case Study of Analytic and Lexical Causatives in Fifteen European Languages. *Folia Linguistica* 50(2).507–542.

Mittmann, Brigitta. 2006. With a Little Help From *Friends* (and others): Lexico-pragmatic Characteristics of Original and Dubbed Film Dialogue. In Christoph Houswitschka, Gabriele Knappe and Anja Müller (eds.), *Anglistentag 2005 Bamberg. Proceedings of the Conference of the German Association of University Teachers of English,* 573–585. Trier: Wissenschaftlicher Verlag Trier.

Molinelli, Piera. 2015. Polite Forms and Sociolinguistic Dynamics in Contacts Between Varieties of Italian. In Carlo Consani (ed.), *Contatto interlinguistico fra presente e passato*, 283–314. Milan: LED.

Odber de Baubeta, Patricia Anne. 1992. Modes of Address: Translation Strategies of the Black Hole. *Ilha do Desterro* 28. 87–107.

Quaglio, Paulo. 2008. Television Dialogue and Natural Conversation: Linguistic Similarities and Functional Differences. In Annelie Ädel and Randi Reppen (eds.), *Corpora and Discourse: The Challenges of Different Settings*, 189–210. Amsterdam: John Benjamins.

R Core Team. 2015. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org/ [Accessed: 12 January 2015].

Schmid, Hans-Jörg. 2014. Lexico-grammatical Patterns, Pragmatic Associations and Discourse Frequency. In Thomas Herbst, Hans-Jörg Schmid and Susen Faulhaber (eds.), *Constructions, Collocations*, *Patterns*, 239–295. Berlin/Boston: de Gruyter Mouton.

Strobl, Carolin et al. 2008. Conditional Variable Importance for Random Forests. *BMC Bioinformatics* 9: 307. Available from: http://www.biomedcentral.com/1471-2105/9/307 [Accessed: 17 March 2015].

Szarkowska, Agnieszka. 2013. *Forms of Address in Polish-English Subtitling*. Frankfurt am Main: Peter Lang.

Szmrecsanyi, Benedikt et al. 2016. Around the World in Three Alternations: Modeling Syntactic Variation in Varieties of English. *English World-Wide* 37(2). 109–137.

Tagliamonte, Sali and R. Harald Baayen. 2012. Models, Forests and Trees of York English: Was/were Variation as a Case Study for Statistical Practice. *Language Variation and Change* 24(2). 135–178.

Warren, Jane. 2006. Address Pronouns in French: Variation Within and Outside the Workplace. *Australian Review of Applied Linguistics* 29(2). 16.1–16.17.

Weiner, E. Judith and William Labov. 1983. Constraints on the Agentless Passive. *Journal of Linguistics* 19. 29–58.

Wierzbicka, Anna. 1985. Different Cultures, Different Languages, Different Speech Ccts: Polish vs. English. *Journal of Pragmatics* 9. 145–178.

Yli-Vakkuri, Valms. 2005. Politeness in Finland: Evasion at all Costs. In Leo Hickey and Miranda Stewart (eds.), *Politeness in Europe*, 189–202. Clevedon: Multilingual Matters.

List of abbreviations

IMP	imperative
IPF	imperfective
PRS	present
2PL	2nd person plural
2SG	2nd person singular